Applied Mathematical Sciences, Vol. 7, 2013, no. 13, 641 - 644

On Multiple Occurrences Shortest Common Superstring Problem

Anna Gorbenko

Department of Intelligent Systems and Robotics
Ural Federal University
620083 Ekaterinburg, Russia
gorbenko.ann@gmail.com

Vladimir Popov

Department of Intelligent Systems and Robotics
Ural Federal University
620083 Ekaterinburg, Russia
Vladimir.Popov@usu.ru

Abstract

In this paper, we consider multiple occurrences shortest common superstring problem. In particular, we consider an approach to solve the problem. This approach is based on an explicit reduction from the problem to the satisfiability problem.

Keywords: multiple occurrences shortest common superstring problem, satisfiability problem, **NP**-complete

Usage of different regularities (see e.g. [1] – [8]) allows us to create systems of robot self-awareness (see e.g. [9] – [17]). In particular, the multiple occurrences shortest common superstring problem (MOSCS) was proposed in [18]. The problem allows us to find regularities with elements of novelty.

Let $\mathcal{S} = \{S_1, S_2, \ldots, S_k\}$ be the set of input words, $S_i \in \Sigma^*$, $1 \leq i \leq k$. Let $|\mathcal{S}|$ denote the total length of all words in $\mathcal{S}$. Let $\#occ(U, V)$ be the number of occurrences (as a factor) of the word U in the word V. The decision version of the MOSCS problem (MOSCS-D) can be formulated as following:

INSTANCE: *A fixed alphabet Σ, a positive integers k and r, a set of input words $\mathcal{S} = \{S_1, S_2, \ldots, S_k\}$, nonnegative integers $m_1, m_2, \ldots, m_k, n_1, n_2, \ldots, n_k$ and positive integer m.*

QUESTION: *Is there a word U such that $|U| \leq r$, $m_i \leq \#occ(S_i, U) \leq n_i$ for all $1 \leq i \leq k$ and $\sum_{i=1}^{k} \#occ(S_i, U) \geq m$?*

Since in practice we are not interested in exponential output, we can assume that m_i and n_i are given as sequences of units for all i such that $1 \leq i \leq k$. In this case, MOSCS-D is **NP**-complete [18]. In this paper, we consider reductions from MOSCS-D to SAT and 3SAT assuming that $m_1, m_2, \ldots, m_k$, n_1, $n_2, \ldots, n_k$ and m are constants. Let $\Sigma = \{a_1, a_2, \ldots, a_p\}; \mathcal{S} = \{S_1, S_2, \ldots, S_k\}$;

$$\varphi = \wedge_{1 \leq i \leq r}((\vee_{1 \leq j \leq p} x[i,j]) \wedge (\wedge_{1 \leq j[1] < j[2] \leq p}(\neg x[i,j[1]] \vee \neg x[i,j[2]])));$$

$$\psi[1,i] = \vee_J \wedge_{1 \leq l \leq m_i} y[i,j[l]],$$

$$J = \{j[l] \mid 1 \leq l \leq m_i, 1 \leq j[1] < j[2] < \ldots j[m_i] \leq r + 1 - |S_i|\};$$

$$\psi[2,i] = \wedge_J \vee_{1 \leq l \leq n_i + 1} \neg y[i,j[l]],$$

$$J = \{j[l] \mid 1 \leq l \leq n_i + 1, 1 \leq j[1] < j[2] < \ldots j[n_i + 1] \leq r + 1 - |S_i|\};$$

$$\psi = \wedge_{1 \leq i \leq k}(\psi[1,i] \wedge \psi[2,i]);$$

$$\tau[i,j] = \wedge_{0 \leq l \leq |S_i| - 1, S_i[l+1] = a_q}(\neg y[i,j] \vee x[j+l,q]);$$

$$\tau = \wedge_{1 \leq i \leq k} \wedge_{1 \leq j \leq r+1-|S_i|} \tau[i,j];$$

$$\rho[i,j] = \vee_{0 \leq l \leq |S_i| - 1, S_i[l+1] = a_q}(y[i,j] \vee \neg x[j+l,q]);$$

$$\rho = \wedge_{1 \leq i \leq k} \wedge_{1 \leq j \leq r+1-|S_i|} \rho[i,j];$$

$$\eta = \vee_J \wedge_{1 \leq s \leq m} y[i[l[s]], j[l[s]]],$$

$$J = \{(i[l[s]], j[l[s]]) \mid (i[l[s]], j[l[s]]) \in \{(i,j) \mid 1 \leq i \leq k, 1 \leq j \leq r + 1 - |S_i|\},$$
$$s \neq t \rightarrow (i[l[s]], j[l[s]]) \neq (i[l[t]], j[l[t]]), 1 \leq s \leq m\};$$

$$\xi = \varphi \wedge \psi \wedge \tau \wedge \rho \wedge \eta.$$

In ξ formula φ can be considered as a choice of a common superstring, for all i formula $\psi[1,i] \wedge \psi[2,i]$ can be considered as a choice of positions of word S_i in the common superstring. It is easy to check that there is a word U such that $|U| \leq r$, $m_i \leq \#occ(S_i, U) \leq n_i$ for all $1 \leq i \leq k$ and $\sum_{i=1}^{k} \#occ(S_i, U) \geq m$ if and only if ξ is satisfiable. Since $m_1, m_2, \ldots, m_k$, n_1, $n_2, \ldots, n_k$ and m are constants, it is possible to apply laws of distributivity to ψ and η. Using this transformation ξ can be represented in conjunctive normal form ξ'. This gives us a reduction from MOSCS-D to SAT.

Using standard transformations (see e.g. [19]) we can easily obtain an explicit transformation ξ' into ζ such that $\xi' \Leftrightarrow \zeta$ and ζ is a 3-CNF. It is clear that ζ gives us an explicit reduction from MOSCS-D to 3SAT.

There is a well known site on which posted solvers for SAT [20]. They are divided into two main classes: stochastic local search algorithms and algorithms improved exhaustive search. All solvers allow the conventional format

for recording DIMACS boolean function in conjunctive normal form and solve the corresponding problem. In addition to the solvers the site also represented a large set of test problems in the format of DIMACS. This set includes a randomly generated problems of 3SAT. We create a generator of natural instances for MOSCS-D. Also we use test problems from [20]. We use algorithms from [20]: fgrasp and posit. Also we design our own genetic algorithm (OA) for SAT which based on algorithms from [20]. Each test was run on a cluster of at least 100 nodes. Selected experimental results are given in Table 1.

time	brute force	fgrasp	posit	OA
average	27.2 h	32 min	36 min	2.8 min
max	29.1 h	15.16 h	11.37 h	1.26 h
best	26.4 h	16.1 min	11.12 min	80 sec

Table 1: Experimental results for 3SAT

ACKNOWLEDGEMENTS. The work was partially supported by Analytical Departmental Program "Developing the scientific potential of high school" 8.1616.2011.

References

[1] A. Gorbenko and V. Popov, The Far From Most String Problem, *Applied Mathematical Sciences*, 6 (2012), 6719-6724.

[2] A. Gorbenko and V. Popov, On the Longest Common Subsequence Problem, *Applied Mathematical Sciences*, 6 (2012), 5781-5787.

[3] A. Gorbenko and V. Popov, The Longest Common Parameterized Subsequence Problem, *Applied Mathematical Sciences*, 6 (2012), 2851-2855.

[4] A. Gorbenko and V. Popov, The set of parameterized k-covers problem, *Theoretical Computer Science*, 423 (2012), 19-24.

[5] V. Yu. Popov, Computational complexity of problems related to DNA sequencing by hybridization, *Doklady Mathematics*, 72 (2005), 642-644.

[6] V. Popov, The approximate period problem for DNA alphabet, *Theoretical Computer Science*, 304 (2003), 443-447.

[7] V. Popov, The Approximate Period Problem, *IAENG International Journal of Computer Science*, 36 (2009), 268-274.

[8] V. Popov, Multiple genome rearrangement by swaps and by element duplications, *Theoretical Computer Science*, 385 (2007), 115-126.

[9] A. Gorbenko and V. Popov, Anticipation in Simple Robot Navigation and Finding Regularities, *Applied Mathematical Sciences*, 6 (2012), 6577-6581.

[10] A. Gorbenko and V. Popov, Robot's Actions and Automatic Generation of Distance Functions for Sequences of Images, *Advanced Studies in Theoretical Physics*, 6 (2012), 1247-1251.

[11] A. Gorbenko and V. Popov, Robot Self-Awareness: Usage of Co-training for Distance Functions for Sequences of Images, *Advanced Studies in Theoretical Physics*, 6 (2012), 1243-1246.

[12] A. Gorbenko and V. Popov, Robot Self-Awareness: Formulation of Hypotheses Based on the Discovered Regularities, *Applied Mathematical Sciences*, 6 (2012), 6583-6585.

[13] A. Gorbenko and V. Popov, The c-Fragment Longest Arc-Preserving Common Subsequence Problem, *IAENG International Journal of Computer Science*, 39 (2012), 231-238.

[14] A. Gorbenko and V. Popov, Robot Self-Awareness: Occam's Razor for Fluents, *International Journal of Mathematical Analysis*, 6 (2012), 1453-1455.

[15] A. Gorbenko and V. Popov, The Force Law Design of Artificial Physics Optimization for Robot Anticipation of Motion, *Advanced Studies in Theoretical Physics*, 6 (2012), 625-628.

[16] A. Gorbenko, V. Popov, and A. Sheka, Robot Self-Awareness: Exploration of Internal States, *Applied Mathematical Sciences*, 6 (2012), 675-688.

[17] A. Gorbenko, V. Popov, and A. Sheka, Robot Self-Awareness: Temporal Relation Based Data Mining, *Engineering Letters*, 19 (2011), 169-178.

[18] A. Gorbenko and V. Popov, Multiple Occurrences Shortest Common Superstring Problem, *Applied Mathematical Sciences*, 6 (2012), 6573-6576.

[19] A. Gorbenko and V. Popov, Task-resource Scheduling Problem, *International Journal of Automation and Computing*, 9 (2012), 429-441.

[20] http://people.cs.ubc.ca/∼hoos/SATLIB/index-ubc.html

Received: November 1, 2012